

**P-CHANNEL DYNAMIC FLASH MEMORY CELLS  
WITH ULTRATHIN TUNNEL OXIDES**

**Related Applications**

5        This application is a Divisional of U.S. Application No. 09/514,627, filed  
February 28, 2000, <sup>NOW U.S. PAT. NO. 6,384,448, ISSUED ON 5-7-2002</sup> which is incorporated herein by reference.

         This application is related to the following co-pending, commonly assigned  
U.S. patent application: "Dynamic Flash Memory Cells with UltraThin Tunnel  
Oxides," attorney docket no. 303.682US2, serial number 09/882,920, filed June 15,  
10    2001.

**Field of the Invention**

         The present invention relates generally to integrated circuit technology and,  
more particularly, to structures and methods involving p-channel flash memory cells  
15    with ultrathin tunnel oxides.

**Background of the Invention**

         The use of the one device cell, invented by Dennard in 1967 (see generally,  
US Patent 3, 387,286, issued to R. H. Dennard on June 4, 1968, entitled "Field  
20    Effect Transistor memory"), revolutionized the computer industry, by significantly  
reducing the complexity of semiconductor memory. This enabled the cost, of what  
was then a scarce commodity, to be drastically reduced.

         Today, dynamic random access memories (DRAMs) are a mainstay in the  
semiconductor industry. DRAMs are data storage devices that store data as charge  
25    on a storage capacitor. A DRAM typically includes an array of memory cells. Each  
memory cell includes a storage capacitor and an access transistor for transferring  
charge to and from the storage capacitor. Each memory cell is addressed by a word  
line and accessed by a bit line. The word line controls the access transistor such that  
the access transistor controllably couples and decouples the storage capacitor to and  
30    from the bit line for writing and reading data to and from the memory cell.

Over the course of time what was a very simple device (a planer capacitor and one transistor) has, because of even shrinking dimensions, become a very complex structure, to build. Whether it is the trench capacitor, favored by IBM, or the stacked capacitor, used by much of the rest of the industry, the complexity and difficulty has increased with each generation. Many different proposals have been proposed to supplant this device, but each has fallen short because of either the speed of the write or erase cycle being prohibitively long or the voltage required to accomplish the process too high. One example of the attempt to supplant the traditional DRAM cell is the so-called electrically erasable and programmable read only memory (EEPROM), or more common today, flash memory.

Electrically erasable and programmable read only memories (EEPROMs) provide nonvolatile data storage. EEPROM memory cells typically use field-effect transistors (FETs) having an electrically isolated (floating) gate that affects conduction between source and drain regions of the FET. A gate dielectric is interposed between the floating gate and an underlying channel region between source and drain regions. A control gate is provided adjacent to the floating gate, separated therefrom by an intergate dielectric.

In such memory cells, data is represented by charge stored on the polysilicon floating gates. The charge is placed on the floating gate during a write operation using a technique such as hot electron injection or Fowler-Nordheim (FN) tunneling. Fowler-Nordheim tunneling is typically used to remove charge from the polysilicon floating gate during an erase operation. A flash EEPROM cell has the potential to be smaller and simpler than a DRAM memory cell. One of the limitations to shrinking a flash EEPROM memory cell has been the requirement for a silicon dioxide gate insulator thickness of approximately 10 nm between the floating polysilicon gate and the silicon substrate forming the channel of a flash field effect transistor. This gate thickness is required to prevent excess charge leakage from the floating gate that would reduce data retention time (targeted at approximately 10 years)

Current n-channel flash memories utilize a floating polysilicon gate over a silicon dioxide gate insulator of thickness of the order 100 Å or 10 nm in a field effect transistor. (See generally, B. Dipert et al., IEEE Spectrum, pp. 48-52 (Oct. 1993). This results in a very high barrier energy of around 3.2 eV for electrons

5 between the silicon substrate and gate insulator and between the floating polysilicon gate and silicon oxide gate insulator. This combination of barrier height and oxide thickness results in extremely long retention times even at 250 degrees Celsius. (See generally, C. Papadas et al., IEEE Trans. on Electron Devices, 42, 678-681 (1995)). The simple idea would be that retention times are determined by thermal

10 emission over a 3.2 electron volt (eV) energy barrier, however, these would be extremely long so the current model is that retention is limited by F-N tunneling off of the charged gate. This produces a lower "apparent" activation energy of 1.5 eV which is more likely to be observed. Since the retention time is determined either by thermal excitation of electrons over the barrier or the thermally assisted F-N

15 tunneling of electrons through the oxide, retention times are even longer at room temperature and/or operating temperatures and these memories are for all intensive purposes non-volatile and are also known as non volatile random access memories (NVRAMs). This combination of barrier height and oxide thickness tunnel oxide thickness is not an optimum value in terms of transfer of electrons back and forth

20 between the substrate and floating gate and results in long erase times in flash memories, typically of the order of milliseconds. To compensate for this, a parallel erase operation is performed on a large number of memory cells to effectively reduce the erase time, whence the name "flash" or "flash EEPROM" originated since this effective erase time is much shorter than the erase time in EEPROMs.

25 P-channel flash memory cells, as shown in Figure 1A having gate oxide thicknesses of approximately 100 Å, have been reported (see generally, T. Ohnakado et al., Digest of Int. Electron Devices Meeting, Dec. 10-13, 1995, Washington D. C., pp. 279-282; T. Ohnakado et al., Digest of Int. Electron Devices Meeting, Dec. 8-11, 1996, San Francisco, pp. 181-184; T. Ohnakado et al., Proc. Symposium on

30 VLSI Technology, June 9-11, 1998), Honolulu, HI, pp. 14-15) and disclosed (see

U.S. Patent 5,790,455, issued Aug. 4, 1998, entitled "Low voltage single supply CMOS electrically erasable read-only memory"). These reported and disclosed p-channel flash memory cells which work similar to n-channel flash memory cells in that they utilize hot electron effects to write data on to the floating gate. If the magnitude of the drain voltage in a PMOS transistor is higher than the gate voltage, then the electric field near the drain through the gate oxide will be from the gate (most positive) towards the drain (most negative). This can and will cause hot electrons to be injected into the oxide and collected by the gate as shown in Figure 1B. The mechanisms reported are either channel hot electron injection, CHE, or band-to-band tunneling induced hot electron injection, BTB. The gate current in PMOS transistors (see generally, I. C. Chen et al., IEEE Electron Device Lett., 4:5, 228-230 (1993); and J. Chen et al., Proceedings IEEE Int. SOI Conf., Oct. 1-3, 1991, pp. 8-9) can actually be much higher than the gate current in NMOS transistors (see generally, R. Ghodsi et al., IEEE Electron Device Letters, 19:9, 354-356 (1998)) due to the BTB tunneling. Negatively, higher gate current in the PMOS transistors resulting from this BTB tunneling effect limits the reliability of deep sub-micron CMOS technology, as reported by R. Ghodsi et al. In other words, the reliability of the PMOS array is lowered because of this higher current in the PMOS device.

In addition to these hot electron effects for electron injection and tunneling through the gate oxide, there is of course also the usual Fowler-Nordheim (FN) tunneling and band-to-band tunneling, BTB, which can and has been used to write and erase conventional tunneling mode n-channel flash memory cells. In the case of p-channel devices with heavily doped p-type source and drain regions where there are no electrons available for tunneling in the conduction band, valence band electrons can tunnel through the gate oxides or insulators. (See generally, C. Salm et al., IEEE Electron Device Letters, 19:7, 213-215 (1998)). In this latter case, the energy barrier is higher which makes the tunneling probability lower. To account for this drawback, silicon germanium (SiGe) floating gates have been used to reduce this barrier. Figures 2A and 2B illustrate one of the primary reliability concerns in

CMOS technology and in flash memory technology. This is the concern of electrons being injected into the silicon substrate with energies over 4.7 eV. These electrons can generate electron hole pairs and the "hot holes" can be injected back into the silicon oxide gate insulator. "Hot hole" injection will lead to reduced barriers for electrons, localized high currents and p-channel threshold voltage shifts. Figure 2B illustrates the threshold voltage shift on a high frequency MOS capacitance vs. voltage (C-V) curve. Such threshold voltage shifts are undesirable in flash memory devices. This is one of the reasons n-channel flash memories use source side tunneling erase, to avoid the build up of trapped holes and/or anomalous positive charge in the transistor gate oxide above the channel.

There is a need in the art to develop p-channel floating gate transistors which can be implemented in deep sub-micron CMOS technology devices, i.e. which can replace DRAM cells in CMOS technology. That is, it is desirable to develop p-channel floating gate transistors which are more responsive, providing faster write and erase times than conventional Flash Memory. It is further desirable that such p-channel floating gate transistors are more reliable, e.g. avoid p-channel threshold voltage shifts and achieve source side tunneling erase.

### **Summary of the Invention**

Previous structures and methods have described the use of reduced barriers between the substrate and polysilicon floating gate and gate insulator by using a different gate insulator rather than silicon dioxide; or using a material other than polysilicon for the floating gate to reduce the barrier between the floating gate and gate insulator. Other structures and methods disclosed previously to increase the tunneling current and reduce the erase time have included the use of roughened silicon surfaces under the tunnel oxide to locally increase electric fields.

The structures and methods described in the present invention include the utilization of dynamic, "enhancement mode," p-channel flash memories with an ultrathin tunnel oxide thickness and the operation of the memory cells on a tunnel-tunnel basis. Both the write and erase operations are performed by tunneling. Since

the tunneling current is not only a function of the applied electric fields but is also an exponential function of the tunnel oxide thickness, thinner gate tunnel oxides will reduce the write and erase operations by orders of magnitude. Thinner tunnel oxides will also result in reduced retention times since the thinner oxides will result in

5 increased thermally assisted tunneling of electrons off the floating polysilicon charge storage gate. According to the teachings of the present invention, a p-channel flash memory with thin tunnel oxides will operate on a dynamic basis, the stored data can be refreshed every few seconds as necessary vs. every few milliseconds for a conventional DRAM; the write and erase operations will however

10 now be orders of magnitude faster than a conventional Flash Memory. In this respect then, the p-channel flash memory now operates in a manner equivalent to DRAMs, except there is no longer a requirement for large stacked storage capacitors or deep trench storage capacitors. The large capacitors are unnecessary since the cell now is active in nature and the transistor provides a large gain. The present

15 invention further provides structures and methods for p-channel floating gate transistors which avoid p-channel threshold voltage shifts and achieve source side tunneling erase

According to one embodiment of the present invention, a p-channel memory cell is provided. The p-channel memory cell includes a control gate. A floating

20 gate is separated from the control gate by a dielectric layer. An oxide layer of less than 50 Angstroms (Å) separates the floating gate from a channel region separating a source and a drain region in a substrate. According to the teachings of the present invention, the floating gate is adapted to hold a charge of the order of  $10^{-17}$  Coulombs for at least 1.0 second at 85 degrees Celsius.

25 According to another embodiment of the present invention, a method for operating a p-channel memory cell is provided. The method includes applying a potential of less than 3.0 Volts across a floating gate oxide which is less than 50 Angstroms, in order to add or remove a charge from a floating gate. The method further includes reading the p-channel memory cell by applying a potential to a

30 control gate of the p-channel memory cell of less than 1.0 Volt.

These and other embodiments, aspects, advantages, and features of the present invention will be set forth in part in the description which follows, and in part will become apparent to those skilled in the art by reference to the following description of the invention and referenced drawings or by practice of the invention.

- 5 The aspects, advantages, and features of the invention are realized and attained by means of the instrumentalities, procedures, and combinations particularly pointed out in the appended claims.

### **Brief Description of the Drawings**

- 10 Figure 1A illustrates an embodiment of a prior art p-channel flash memory cell having a floating gate oxide of 100 Angstroms.

Figure 1B illustrates an embodiment of the operation for a prior art p-channel flash memory cell having a floating gate oxide of 100 Angstroms.

- 15 Figure 2A illustrates the negative gate voltage(s) stress on PMOS devices with n-type substrates or wells.

Figure 2B is a graph of capacitance vs. gate voltage illustrating the voltage shift on high frequency, prior art p-channel flash memory cells due to the effect of hot hole injection and/or anomalous positive charge generation.

- 20 Figure 3A illustrates a p-channel flash memory cell having a floating gate oxide of less than 50 Angstroms according to the teachings of the present invention.

Figure 3B is an energy band diagram which illustrates generally the write operation and charge injection to the floating gate of the p-channel flash memory cell of present invention from the substrate and/or n-well.

- 25 Figure 3C is an energy band diagram which illustrates generally the write operation and charge injection to the floating gate of the p-channel flash memory cell of present invention from the p<sup>+</sup> source region.

Figure 4A illustrates a p-channel flash memory cell having a floating gate oxide of less than 50 Angstroms according to the teachings of the present invention.

Figure 4B is an energy band diagram which illustrates generally the erase operation and charge tunneling from the floating gate to the p+ source region for the p-channel flash memory cell of present invention.

Figure 5A and 5B further illustrate the erase operation and charge tunneling from the floating gate to the p+ source region for the p-channel flash memory cell of the present invention.

Figures 6A and 6B are graphs that illustrate generally the temperature dependence and the dependence of the retention time on gate tunneling oxide thickness based on an extension of published data.

Figure 7 is a graph that illustrates generally the current density through ultrathin oxides based on an extension of published data from which an erase time can be calculated.

Figure 8 is a graph that illustrates generally the mean time to failure of the ultra thin tunnel oxides for p-channel flash memory cells according to the teachings of the present invention.

Figure 9 illustrates an embodiment of a p-channel flash memory cell coupled to a sense amplifier according to the teachings of the present invention.

Figure 10 is a block diagram of a system according to one embodiment of the present invention.

Figure 11 illustrates an embodiment of a memory array according to the teachings of the present invention, as can be included in a memory device.

### **Detailed Description of the Invention**

In the following detailed description of the invention, reference is made to the accompanying drawings which form a part hereof, and in which is shown, by way of illustration, specific embodiments in which the invention may be practiced. In the drawings, like numerals describe substantially similar components throughout the several views. These embodiments are described in sufficient detail to enable those skilled in the art to practice the invention. Other embodiments may be utilized and structural, logical, and electrical changes may be made without departing from



the scope of the present invention. The terms wafer and substrate used in the following description include any structure having an exposed surface with which to form the integrated circuit (IC) structure of the invention. The term substrate is understood to include semiconductor wafers. The terms wafer and substrate used in the following description include any base semiconductor structure. Both are to be understood as including bulk silicon material, silicon-on-sapphire (SOS) technology, silicon-on-insulator (SOI) technology, thin film transistor (TFT) technology, doped and undoped semiconductors, epitaxial layers of silicon supported by a base semiconductor structure, as well as other semiconductor structures well known to one skilled in the art. Furthermore, when reference is made to a wafer or substrate in the following description, previous process steps may have been utilized to form regions/junctions in the base semiconductor structure and layer formed above, and the terms wafer or substrate include the underlying layers containing such regions/junctions and layers that may have been formed above. The term conductor is understood to include semiconductors, and the term insulator is defined to include any material that is less electrically conductive than the materials referred to as conductors. The following detailed description is, therefore, not to be taken in a limiting sense, and the scope of the present invention is defined only by the appended claims, along with the full scope of equivalents to which such claims are entitled.

The technique described here is the utilization of p-channel flash memories with an ultrathin tunnel oxide thickness and the operation of the memory cells on a tunnel-tunnel basis. Both the write and erase operations are performed by tunneling. Since the tunneling current is not only a function of the applied electric fields but is also an exponential function of the tunnel oxide thickness, thinner gate tunnel oxides will reduce the write and erase operations by orders of magnitude. Thinner tunnel oxides will also result in reduced retention times since the thinner oxides will result in increased thermally assisted tunneling of electrons off the floating polysilicon charge storage gate. A flash memory with thin tunnel oxides will operate on a dynamic basis, the stored data can be refreshed every few seconds as necessary; the

write and erase operations will however now be orders of magnitude faster. In this respect then, the device now operates in a manner equivalent to DRAMs, except there is no longer a requirement for large stacked storage capacitors or deep trench storage capacitors. The large capacitors are unnecessary since the cell now is active in nature and the transistor provides a large gain.

Figure 3A illustrates a p-channel flash memory cell 301, or p-channel static device 301, having a floating gate oxide of less than 50 Angstroms according to the teachings of the present invention. The p-channel flash memory cell 301 includes a p-channel transistor. As shown in Figure 3A, the p-channel flash memory cell 301 of the present invention includes a control gate 302 and a floating gate 304. The control gate 302 is separated from the floating gate 304 by an intergate dielectric layer, or intergate oxide 303. In one embodiment the intergate dielectric layer 303 includes a layer of silicon dioxide ( $\text{SiO}_2$ ). In an alternative embodiment, the intergate dielectric layer 303 includes a silicon nitride layer ( $\text{Si}_3\text{N}_4$ ) or any other suitable dielectric layer 303, the invention is not so limited.

Figure 3A illustrates that the floating gate 304 is separated by an oxide layer, or tunnel gate oxide 305, from a channel region 306. The channel region separates a source region 310 from a drain region 312 in a substrate 300. As one of ordinary skill in the art will understand upon reading this disclosure the source and drain regions, 310 and 312 respectively, include p-type doped source and drain regions, 310 and 312. The p-type doped source and drain regions, 310 and 312, can include heavily doped (p+) source and drain regions, 310 and 312. The substrate includes an n-type substrate or n-well 300.

According to the teachings of the present invention, the oxide layer 305 is an oxide layer which has a thickness ( $t_1$ ) of less than 50 Angstroms ( $\text{\AA}$ ). In one embodiment of the present invention, the oxide layer is approximately 30 Angstroms ( $\text{\AA}$ ). In one exemplary embodiment of the present invention, the oxide layer is approximately 23 Angstroms ( $\text{\AA}$ ).

Figure 3B is an energy band diagram which illustrates generally the write operation and charge injection from the n-type substrate or n-well 300 to the floating

gate 304, as shown in Figure 3A, when a positive potential is applied to the floating gate 304. Figure 3C is an energy band diagram which illustrates generally the write operation and charge injection to the floating gate 304 in the PMOS flash memory cell 301 of Figure 3A from the p+ source region 310 when a positive potential is applied to the floating gate 304. Injection from the n-type substrate 300 is the preferred technique since the barrier is lower than for injection from the valence band of the p+ source region as disclosed by C. Salm et al., IEEE Electron Device Letters, vol. 19, no. 7, 213-215 (July 1998).

Figure 4A again shows the p-channel flash memory cell 401, or p-channel static device 401, having a floating gate oxide of less than 50 Angstroms according to the teachings of the present invention as shown in Figure 3A. Figure 4A is provided to illustrates the erase operation and tunneling of charge from the floating gate 404 to the source 410, when the floating gate is driven to a negative potential by pulsing the control gate 402 to a negative voltage. Figure 4B is an energy band diagram which illustrates generally the erase operation and tunneling of charge from the floating gate 404 to the source 410, when the floating gate is driven to a negative potential by pulsing the control gate 402 to a negative voltage. Tunneling to the p+ source 410 is the preferred technique rather than to the n-type substrate or n-well 400 to avoid charge buildup in the tunnel gate oxide 405 over the transistor channel 406.

Figure 5A again shows the p-channel flash memory cell 501, or p-channel static device 501, having a floating gate oxide of less than 50 Angstroms according to the teachings of the present invention. As shown in Figure 5A the floating gate 504 has a bottom surface area 509 in contact with the oxide layer, or floating gate oxide 505. In one embodiment according to the teachings of the present invention, the bottom surface area has an area of approximately  $10^{-10}$  cm<sup>2</sup>. Figure 5A illustrates the source side tunneling for F-N tunneling erase or thermally assisted tunneling charge leakage of stored electrons off of a polysilicon floating gate 504. In one embodiment, the polysilicon floating gate 504 is an n-type polysilicon floating gate 504, such as an n+ polysilicon floating gate 504. In an alternative embodiment, the

polysilicon floating gate 504 is an p-type polysilicon floating gate 504, such as an p+ polysilicon floating gate 504. In another embodiment, the polysilicon floating gate 504 is a polysilicon-germanium floating gate 504. According to the teachings of the present invention, an n-type polysilicon floating gate 504, such as an n+ polysilicon floating gate 504 is preferred rather than a p-type polysilicon or p-type polysilicon-germanium floating gate since the tunneling barrier for electrons is lower. Figure 5B is an energy band diagram which illustrates generally the thermally assisted tunneling or tunneling erase of charge from the floating gate 504 to the source 510, when the floating gate is driven to a negative potential by pulsing the control gate 502 to a negative voltage.

Figures 6A and 6B provide an estimate of the temperature dependence and the dependence of the retention time on gate tunneling oxide thickness based on an extension of published data as disclosed in C. Papadas et al., IEEE Trans. on Electron Devices, vol. 42, 678-681 (April 1995). In Figure 6A the X-axis represents normalized temperature (1000/T) and the Y-axis represents the retention time (tr) relative values in log base 10 (log10). Thus, in Figure 6A temperature T1 is approximately 250 degrees Celsius, temperature T2 is approximately 85 degrees Celsius, and temperature T3 is approximately 27 degrees Celsius. From this published data the temperature dependance of the retention time (tr) can be given by:

$$tr \propto e^{(\Delta E/kT)}$$

where  $\Delta E$  is the activation energy of the process in the gate oxide, k is the Boltzmann constant, and T is temperature. Using the graph shown in Figure 6A and the above formula, the change in the order of magnitude of the retention time (tr) for the various temperatures shown can be determined.

Figure 6B is a graph of threshold voltage (VT) vs retention time (tr) in hours calculated at a temperature of approximately 250 degrees Celsius. The graph illustrates the retention time (tr) as a function of threshold voltage (VT) for a number of different tunnel gate oxide thicknesses (TH). In Figure 6B, TH1

represents a gate oxide thickness of approximately 60 Angstroms (Å), TH2 represents a gate oxide thickness of approximately 70 Angstroms (Å), TH3 represents a gate oxide thickness of approximately 80 Angstroms (Å), TH4 represents a gate oxide thickness of approximately 90 Angstroms (Å). Further

5 calculations can be made based on an extension of the published data represented in Figures 6A and 6B.

From Figure 6B, the retention time (tr), at a temperature of 250 degrees Celsius and an applied potential of 2.5 Volts across the gate oxide, can be expressed as a function of the tunnel gate oxide thickness (tox) by the formula:

10

$$tr \propto 10^{18.65} \times e^{(-2000/tox \text{ (Å)})}$$

where  $10^{18.65}$  and  $e^{(-2000/tox \text{ (Å)})}$  are constants derived to fit the data for the various gate oxide thicknesses shown on the graph. Using this formula, the retention time (tr)

15 can be calculated for a number of different gate oxide thicknesses. Thus, for a gate oxide having a thickness of 3 nm or 30 Å, when the temperature is 250 at degrees Celsius and a potential of 2.5 Volts is applied across the gate oxide, the retention time (tr) is approximately  $10^{-10}$  hours. From an extension of the published data shown in Figures 6A and 6B and using the formulas for tr derived from these

20 graphs, tr can similarly be calculated when a potential of 2.5 Volts is applied across a 30 Å thick gate oxide at other temperatures. That is, a value for tr was calculated when a potential of 2.5 Volts is applied across a 30 Å thick gate at a temperature of 250 degrees Celsius. The change in the order of magnitude of the retention time (tr) for different temperatures is provided by the formula for tr derived from Figure 6A.

25 Multiplying the appropriate change in the order of magnitude for tr from a temperature of 250 degrees Celsius to a different temperature on the graph of Figure 6A will give the retention time (tr) at that different temperature. Thus, the retention time (tr) when a potential of 2.5 Volts is applied across a 30 Å thick gate oxide at 250 degrees Celsius can be extrapolated to provide the retention time when only the

30 temperature is varied. These calculations show a retention time (tr) of  $10^{-3}$  hours, or

a few seconds, at a temperature of 85 degrees Celsius when a potential of 2.5 Volts is applied across a 30 Å thick gate oxide. These calculations further show a retention time (tr) of 10 hours at a temperature of 27 degrees Celsius, or approximately room temperature.

5           These calculations show a retention time (tr) comparable to or longer than that of current DRAMs. These estimates are approximate since the data is based on results for much thicker oxides, however, 2.5 V is far in excess of the operating gate voltages of more like less than 1.0 V used on ultrathin gate oxides. The actual retention times are likely to be much longer at these lower operating voltages. Thus, 10 flash memories with 30 Å gate oxides may no longer be non-volatile but their retention times will be long in comparison to DRAMs. Further calculations can be performed in a similar manner to arrives at retention times for ultra thin gate oxides of other thicknesses, e.g. 23 Å.

Another technique to estimate the retention time (tr) can be taken from the 15 published curves for current densities which are less than  $5 \times 10^{-7}$  A/cm<sup>2</sup> for 30 Å gate oxides with voltages across the dielectric of much less than 1.0 Volt as disclosed by T. P. Ma et al., IEEE Electron Device Letters, vol. 19: no. 10, pp. 388-390 (1998). If the charge stored on the floating gate is of the order of  $10^{-7}$  Coulombs/cm<sup>2</sup>, then the retention time is of the order of seconds. There are some 20 inconsistencies in the literature provided here. That is, the tunneling current shown in Figures 5A and 5B is F-N or band-to-band tunneling current, the same as disclosed by T. P. Ma et al., IEEE Electron Device Letters, vol. 19: no. 10, pp. 388-390 (1998), which is normally assumed to be independent of temperature. Whereas, the tunneling current represented in Figures 6A and 6B, as disclosed by C. Papadas 25 et al., IEEE Trans. on Electron Devices, vol. 42, pp. 678-681 (April 1995), is thermally assisted tunneling which is a strong function of temperature. However, either technique gives the same estimate for retention times (tr) on the order of seconds at 85 degrees Celsius.

In addition to the long retention time in comparison to DRAMs, the thinner 30 tunneling oxides, according to the teachings of the present invention, will also result

in much shorter write and erase times than is customary in conventional flash memories. Figure 7 is a graph that illustrates generally the current density through ultrathin oxides of various thicknesses, e.g. TH<sub>5</sub> and TH<sub>6</sub>, as a function of the voltage, or potential, applied across the gate oxide, based on an extension of above published data. The graph shown in Figure 7 can be used to calculate an erase time of the present invention. In one embodiment of the present invention, a p-channel flash memory cell is formed with a 3 nm or 30 Å tunnel gate oxide. Based on the published data in Figure 7, a 3 nm or 30 Å gate oxide, TH<sub>6</sub>, with a voltage of -3.0 V across the oxide, and an electric field of 10 megavolts per centimeter (MV/cm) will have a current density of about  $5 \times 10^{-3}$  Amperes/cm<sup>2</sup>. If this current flows for 20 microseconds (μsec) to write or erase, then the stored charge on the floating gate will change by 10<sup>-7</sup> coulombs/cm<sup>2</sup>. In the case of erase with -3.0 V applied to the floating gate and a gate area of 10<sup>-10</sup> cm<sup>2</sup>, the magnitude of the change of charge on the floating gate will be 10<sup>-17</sup> coulombs or about 100 electrons. This change in charge will result in a change of potential of the floating gate of about 80 milliVolts (mV) since the gate capacitance is about 0.13 femto Farads (fF). Because this device works on a tunnel-tunnel basis, the write time will similarly be 20 microseconds (μsec) at +3.0 V on the gate.

In another embodiment of the present invention, a p-channel flash memory cell is formed with a 2.3 nm or 23 Å tunnel gate oxide. Based on the above published data, if a 2.3 nm or 23 Å gate oxide, TH<sub>5</sub>, is used with a voltage of -2.3 V the current density will be orders of magnitude larger, e.g. 0.5 A/cm<sup>2</sup>, as shown in Figure 7. In the case of erase with -2.3 V applied to the floating gate and a gate area of 10<sup>-10</sup> cm<sup>2</sup>, the time necessary to effect a change of charge on the floating gate of 10<sup>-17</sup> coulombs, or about 100 electrons, will be reduced to 200 nanoseconds (nS). Again, since the device of the present invention works on a tunnel-tunnel basis, the write time will similarly be 200 nanoseconds (nS) at +2.3 V on the gate. This will reduce the write and erase times to 200 nanoseconds (nS).

Figure 8 is a graph that illustrates generally the mean time to failure in minutes vs. the electric field applied across ultra thin tunnel oxides as disclosed by

G. Lucovsky et al., IEEE Electron Device Letters, vol. 20: no. 6, pp. 262-264 (1999). This recent data suggests that the meantime to failure of ultrathin gate oxides and the total charge to breakdown, QBD, of ultrathin gate oxides may be orders of magnitude larger than tunnel oxides in the 50 Å to 100 Å thickness range.

5 This may result since an electron energy of greater than 1.7 eV and a total potential difference of greater than the order of 4.7 volts is required to cause damage to the oxide as disclosed by G. Lucovsky et al., IEEE Electron Device Letters, vol. 20: no. 6, pp. 262-264 (1999); and N. Patel et al., Appl. Phys. Lett., vol. 64: no. 14, 1809-11 (1994). In other words, there is likely a threshold electron energy before significant

10 damaging can occur. From the graph shown in Figure 8, an electric field of 10 MV/cm across a gate oxide having a thickness of 3 nm ( $TH_7$ ) has a mean time to failure ( $t_1$ ) of approximately  $10^7$  minutes, or 20 years. If each operation has a duration of 20 μsec, the tunnel gate oxide can perform  $30 \times 10^{12}$  operations. As shown above, an electric field of 10 MV in a 3 nm tunnel gate oxide in a p-channel

15 flash memory cell of the present invention results in a current density of 5 mA/cm<sup>2</sup>. A current density of 5 mA/cm<sup>2</sup> for a duration of 20 μsec, repeated for a total of  $30 \times 10^{12}$  operations, results in a total charge transferred in one direction through the tunnel gate oxide of  $3 \times 10^6$  Coulombs/cm<sup>2</sup>. Thus, in one embodiment of the present invention, the p-channel flash memory cells having 3.0 V operating voltages and 30

20 Å (3.0 nm), tunnel oxides will likely have a meantime to failure of 20 years and a total charge to failure, QBD, of  $3 \times 10^6$  Coulombs/cm<sup>2</sup>.

Further, the graph shown in Figure 8, as disclosed by G. Lucovsky et al., IEEE Electron Device Letters, vol. 20: no. 6, pp. 262-264 (1999), illustrates that an electric field of 10 MV/cm across a gate oxide having a thickness of 2.3 nm ( $TH_8$ )

25 should have an approximate mean time to failure ( $t_2$ ) of approximately 1000 years. For a current a current density of 5 mA/cm<sup>2</sup>, this would result in a total charge to failure QBD of  $1.5 \times 10^{10}$  Coulombs/cm<sup>2</sup>. In another embodiment of the present invention, a p-channel flash memory cell having a tunnel gate oxide of 2.3 is used. According to the teachings of the present invention, a p-channel flash memory cell

30 having a tunnel gate oxide of 2.3 nm will have a write and erase operation duration



of 200 nanoseconds (ns). If the device of the present invention performs a total of  $10^{15}$  operations, each with a current density of  $0.5 \text{ A/cm}^2$  and a duration 200 ns, then the total charge transferred is  $10^8 \text{ Coulombs/cm}^2$ . This is far less than the total charge to failure, QBD, of  $1.5 \times 10^{10} \text{ Coulombs/cm}^2$ , derived from the graph in

5 Figure 8.

Thus, the p-channel flash memory cells formed according to the teachings of the present invention, having ultrathin gate oxides operated with low voltages, suffer reduced damage and have very large total charge to failure. The p-channel flash memory cells of the present invention are adapted to have a reliability of a number  
10 of cycles of performance of approximately  $10^{12}$  to  $10^{15}$  cycles over a lifetime of the device depending on the tunnel gate oxide thickness, e.g.  $30 \text{ \AA}$  or  $23 \text{ \AA}$ . The p-channel flash memory cells of the present invention, which can perform a total of  $10^{12}$  to  $10^{15}$  operations, offer increased endurance over that of a conventional non volatile flash memory cell. It has also been shown that a p-channel flash memory  
15 cell, or p-channel static device having a p-channel transistor formed according to the teachings of the present invention includes a floating gate which is adapted to hold a charge of the order of  $10^{-17} \text{ Coulombs}$  for longer than 10 hours at 20 degrees Celsius. The floating gate on the p-channel flash memory cell of the present invention is also adapted to hold a charge of the order of  $10^{-17} \text{ Coulombs}$  for at least  
20 1.0 second at 85 degrees Celsius.

Figure 9 illustrates an embodiment of a p-channel flash memory cell 901 coupled to a sense amplifier 902 by a bit line, or digitline 903, according to the teachings of the present invention. Figure 9 is useful in illustrating the read operation of the p-channel flash memory cells formed according to the teachings of  
25 the present invention. According to the teachings of the present invention, the p-channel flash memory cells use an operating voltage of 1.0 Volt or less. That is, a potential of 1.0 Volts or less is applied to the control gate of the p-channel flash memory cell during a read operation. To calculate the effectiveness of the sense operation, a 100 mV change in threshold voltage ( $\Delta VT$ ) for the p-channel flash  
30 memory cell is assumed as a consequence of the change in stored charge, e.g.

approximately  $10^{-17}$  Coulombs, on the floating gate. The charge sensed by the sense amplifier 902 during the read operation can be calculated as follows. The drain to source current is given by:

$$|IDS| = (\epsilon C_0) (W/L) \frac{(VGS - VT)^2}{2}$$

where  $\epsilon$  is the carrier mobility,  $C_0$  is the tunnel gate oxide capacitance, W and L are the width and length of the device respectively, VGS is the potential applied across the floating gate to source region, and VT is the threshold voltage for the device. Since VT approaches  $VT + \Delta VT$ , the equation can be restated as:

$$|\Delta I| = \left( 100 \frac{\mu A}{V^2} \right) (1/1) ((VGS - VT) \Delta VT)$$

substituting in appropriate values for  $\epsilon$  and  $C_0$ , and assuming a device having a 1 to 1 width to length ratio. Taking VGS - VT to be approximately 0.5 Volts and with the change in threshold voltage ( $\Delta VT$ ) to be approximately 0.1 Volts, then the change in current  $\Delta I$  is as follows:

$$\Delta I = 100 \frac{\mu A}{V^2} (1/1) (0.5) (0.1) = 5 \mu A.$$

If the read current is sensed for 1.0 nanosecond,  $t_{sense} = 1$  ns, then the total charge integrated on the bit data sense line 903 will be 5 femto Coulombs (fC).

$$\Delta Q = 5 \times 10^{-15} C$$

$$\Delta Q = 5 \text{ fC}$$

Since the transistor is an active device with gain, there is a charge amplification by a factor of 500, the change in charge on the floating gate is 0.01 fC and the change in charge on the bit line 903 is 5 fC. As one of ordinary skill in the art will understand

upon reading this disclosure, a charge 5 fC is sufficient for a differential sense amplifier 902 to detect as shown in Figure 9.

Figure 10 illustrates a block diagram of an embodiment of a circuit on a single substrate 1001, or an electronic system on a chip 1001 according to the teachings of the present invention. In the embodiment shown in Figure 10, the circuit, or system 1001 includes a memory device 1000 which has an array of memory cells 1002, address decoder 1004, row access circuitry 1006, column access circuitry 1008, control circuitry 1010, and input/output circuit 1012. Also, as shown in Figure 10, the circuit 1001 includes a processor 1014, or memory controller for memory accessing. The memory device 1000 receives control signals from the processor 1014, such as WE\*, RAS\* and CAS\* signals over wiring or metallization lines. The memory device 1000 is used to store data which is accessed via I/O lines. It will be appreciated by those skilled in the art that additional circuitry and control signals can be provided, and that the memory device 1000 has been simplified to help focus on the invention. At least one of the memory cells 1002 has a p-channel transistor, or p-channel flash device, e.g. a p-channel floating gate tunnel oxide (FLOTOX) transistor having an ultrathin gate oxide according to the teachings of the present invention.

It will be understood that the embodiment shown in Figure 10 illustrates an embodiment for the circuitry of a dynamic p-channel flash memory array with ultra thin tunnel oxides according to the teachings of the present invention. The illustration of a circuit or system 1001 as shown in Figure 10 is intended to provide a general understanding of one application for the structure and circuitry of the present invention, and is not intended to serve as a complete description of all the elements and features of a dynamic p-channel flash memory array with ultra thin tunnel oxides. Further, the invention is equally applicable to any size and type of memory circuit 1001 using the dynamic p-channel flash memory cells with ultra thin tunnel oxides of the present invention and is not intended to be limited to the described above. As one of ordinary skill in the art will understand, such single-

package processing units as described in Figure 10 reduce the communication time between the processor and the memory circuit.

As recognized by those skilled in the art, circuits of the type described herein are generally fabricated as an integrated circuit containing a variety of

5 semiconductor devices. The integrated circuit is supported by a substrate. Integrated circuits are typically repeated multiple times on each substrate. The substrate is further processed to separate the integrated circuits into dies as is well known in the art.

Other circuits containing the dynamic p-channel flash memory cells with  
10 ultra thin tunnel oxides described in this disclosure include circuitry for use in memory modules, device drivers, power modules, communication modems, processor modules, and application-specific modules, and may include multilayer, multichip modules. Such circuitry can further be a subcomponent of a variety of electronic systems, such as a clock, a television, a cell phone, a personal computer,  
15 an automobile, an industrial control system, an aircraft, and others.

Figure 11 illustrates an embodiment of a memory array 1100, according to the teachings of the present invention, as can be included in a memory device, e.g. on a memory chip/die. The memory array shown in Figure 11 includes a plurality of memory cells, 1102-0, 1102-1, . . . , 1102-N. The plurality of memory cells, 1102-0,  
20 1102-1, . . . , 1102-N, includes at least one p-channel flash memory cell formed according to the teachings of the present invention. That is, the at least one p-channel flash memory cell includes a source region and a drain region with a channel therebetween. A floating gate is separated from the channel region by an oxide layer, or tunnel gate oxide having a thickness of less than 50 Å. As shown in  
25 Figure 11, the plurality of memory cells are coupled to a plurality, or number of sense amplifiers 1106-0, 1106-1, . . . , 1106-N via a number of bit lines, or digitlines, D0\*, D0, D1\*, D1, . . . , DN\*. Figure 11 is illustrative of the manner in which the dynamic p-channel flash memory cells of the present invention can be used in a folded bit line configuration, in substitution for a conventional folded bit line  
30 DRAM array. One of ordinary skill in the art will understand upon reading this

disclosure, that the p-channel flash memory cells of the present invention can further be used in an open bit line configuration or any other digitline twist scheme. The invention is not so limited. One of ordinary skill in the art will further understand upon reading this disclosure that p-channel flash memory cells of the present

5 invention allow a smaller change in stored charge on the floating gate to be detected.

The Figures presented and described in detail above are similarly useful in describing the methods of the present invention. That is one embodiment of the present invention includes forming a p-channel memory cell. Forming the p-channel memory cell includes forming an oxide layer of less than 50 Angstroms (Å)  
10 on a substrate having a channel region separating a source and a drain region in the substrate. A floating gate is formed on the oxide layer. A dielectric layer is formed on the floating gate. And, a control gate is formed on the dielectric layer. In one embodiment, forming the oxide layer includes forming the oxide layer to have a thickness of 23 Angstroms (Å). According to the teachings of the present invention,  
15 forming a floating gate includes forming a floating gate which is adapted to hold a charge on the order of  $10^{-17}$  Coulombs for longer than 10 hours at 20 degrees Celsius. Also, forming a floating gate includes forming a floating gate which is adapted to hold a charge of the order of  $10^{-17}$  Coulombs for at least 1.0 second at 85 degrees Celsius. According to the teachings of the present invention, forming the  
20 floating gate includes forming a floating gate which has a bottom surface area in contact with the oxide layer of approximately  $10^{-10}$  cm<sup>2</sup>. Forming the p-channel memory cell includes forming the p-channel memory cell to operate at a voltage of approximately 1.0 Volts applied to the control gate.

Another embodiment of the present invention includes a method for forming  
25 a p-channel transistor. The method for forming the p-channel transistor includes forming an oxide layer of less than 50 Angstroms (Å) on a substrate having a channel region separating a source and a drain region in the substrate. A floating gate is formed on the oxide layer. Forming the floating gate includes forming a floating gate which is adapted to hold a charge on the order of  $10^{-17}$  Coulombs for  
30 longer than 1.0 hour at 20 degrees Celsius. In one embodiment, forming the oxide

layer includes forming the oxide layer to have a thickness of 23 Angstroms (Å). Also, forming a floating gate includes forming a floating gate which is adapted to hold a charge of the order of  $10^{-17}$  Coulombs for at least 1.0 second at 85 degrees Celsius. According to the teachings of the present invention, forming the floating gate includes forming a floating gate which has a bottom surface area in contact with the oxide layer of approximately  $10^{-10}$  cm<sup>2</sup>. Forming the p-channel transistor includes forming the p-channel transistor cell to operate at a voltage of approximately 1.0 Volts applied to the control gate.

In one embodiment, forming the p-channel transistor further includes forming an intergate dielectric on the floating gate and forming a control gate on the intergate dielectric. Further, forming the p-channel transistor includes forming the p-channel transistor to have an operating voltage of less than 2.5 Volts across the oxide layer.

A method of forming a memory device is similarly included within the scope of the present invention. According to the teachings of the present invention, forming a memory device includes forming a plurality of memory cells such that forming the plurality of memory cells includes forming at least one p-channel memory cell. Forming the at least one p-channel memory cell includes forming an oxide layer of less than 50 Angstroms (Å) on a substrate having a channel region separating a source and a drain region in the substrate. A floating gate is formed on the oxide layer such that the floating gate is adapted to hold a charge on the order of  $10^{-17}$  Coulombs for longer than 1.0 hour at 20 degrees Celsius. At least one sense amplifier is formed. Forming at least one sense amplifier includes coupling the at least one amplifier to the plurality of memory cells. In one embodiment, forming an oxide layer of less than 50 Angstroms (Å) includes forming the oxide layer to have a thickness of 23 Angstroms (Å). Forming the floating gate further includes forming a floating gate which is adapted to hold a charge on the order of  $10^{-17}$  Coulombs for at least 1.0 second at 85 degrees Celsius.

Forming the memory device according to the teachings of the present invention further includes forming the at least one p-channel memory cell to have an

operating voltage of approximately 1.0 Volt as applied to the control gate in order to perform a read operation on the memory device. Also, forming the memory device according to the teachings of the present invention further includes forming the at least one p-channel memory cell to have an operating voltage of less than 2.5 Volts as applied across the oxide layer in order to perform a write and/or erase operation on the memory device.

Another embodiment of the present invention includes a method for operating a p-channel memory cell. The method includes applying a potential of less than 3.0 Volts across a floating gate oxide, or tunnel gate oxide layer which is less than 50 Angstroms, in order to add or remove a charge from a floating gate. The floating gate is adapted to hold a charge of the order of  $10^{-17}$  Coulombs for longer than 10 hours at 20 degrees Celsius and is also adapted to hold a charge of the order of  $10^{-17}$  Coulombs for at least 1.0 second at 85 degrees Celsius. The method also includes reading the p-channel memory cell by applying a potential to a control gate of the p-channel memory cell of less than 1.0 Volt. In one embodiment of this method, applying a potential of less than 3.0 Volts across a floating gate oxide in order to add or remove a charge from a floating gate includes applying the potential for less than 20 microseconds. The method of the present invention further includes refreshing the p-channel memory cell to renew a charge on the floating gate at 1.0 second intervals. In one embodiment, refreshing the p-channel memory cell to renew a charge on the floating gate at second 1.0 intervals includes renewing a charge of approximately 100 electrons on the floating gate.

In another embodiment of the method for operating a p-channel memory cell, the method includes refreshing a charge on a floating gate of the p-channel memory cell by applying an electric field of approximately 10 mega Volts/centimeter (MV/cm) across a floating gate oxide. The floating gate is adapted to hold a charge of the order of  $10^{-17}$  Coulombs for longer than 10 hours at 20 degrees Celsius and is also adapted to hold a charge of the order of  $10^{-17}$  Coulombs for at least 1.0 second at 85 degrees Celsius. According to the teachings of the present invention the floating gate oxide is less than 30 Angstroms. In this

embodiment, the method further includes reading the p-channel memory cell by applying a potential to a control gate of the p-channel memory cell of less than 1.0 Volt. According to this embodiment, refreshing a charge on a floating gate for the p-channel memory cell by applying an electric field of approximately 10 mega  
5 Volts/centimeter (MV/cm) across a floating gate oxide of less than 30 Angstroms includes restoring a charge of approximately  $10^{-17}$  Coulombs to the floating gate. Also, according to this embodiment, refreshing a charge on a floating gate for the p-channel memory cell by applying an electric field of approximately 10 mega Volts/centimeter (MV/cm) across a floating gate oxide of less than 30 Angstroms  
10 includes applying the electric field for less than 20 microseconds. Further, the method includes refreshing the p-channel memory cell to renew a charge on the floating gate at 1.0 second intervals.

Another embodiment of the present invention includes a method for operating a p-channel memory cell. Here, the method includes applying a potential  
15 of approximately 2.3 Volts across a floating gate oxide in order to add or remove a charge from a floating gate. In this embodiment, applying a potential of approximately 2.3 Volts across a floating gate oxide includes applying a potential of approximately 2.3 Volts across a floating gate oxide which has a thickness of approximately 23 Angstroms. The floating gate is adapted to hold a charge of the  
20 order of  $10^{-17}$  Coulombs for longer than 10 hours at 20 degrees Celsius and is also adapted to hold a charge of the order of  $10^{-17}$  Coulombs for at least 1.0 second at 85 degrees Celsius. The method further includes reading the p-channel memory cell by applying a potential to a control gate of the p-channel memory cell of less than 1.0 Volt. In this embodiment, applying a potential of approximately 2.3 Volts across a  
25 floating gate oxide includes applying the potential for less than 200 nanoseconds. In this embodiment, the method further includes refreshing the p-channel memory cell to renew a charge on the floating gate at 1.0 second intervals. Further, in this embodiment, refreshing the p-channel memory cell to renew a charge on the floating gate at 1.0 second intervals includes renewing a charge of approximately 100  
30 electrons on the floating gate.



### Conclusion

Thus, structures and methods for dynamic p-channel flash memory cells with ultra thin tunnel oxides, have been shown. Both the write and erase operations are performed by tunneling. Since the tunneling current is not only a function of the applied electric fields but is also an exponential function of the tunnel oxide thickness, thinner gate tunnel oxides will reduce the write and erase operations by orders of magnitude. Thinner tunnel oxides will also result in reduced retention times since the thinner oxides will result in increased thermally assisted tunneling of electrons off the floating polysilicon charge storage gate. According to the teachings of the present invention, a p-channel flash memory with thin tunnel oxides will operate on a dynamic basis, the stored data can be refreshed every few seconds as necessary; the write and erase operations will however now be orders of magnitude faster. In this respect then, the p-channel flash memory now operates in a manner equivalent to DRAMs, except there is no longer a requirement for large stacked storage capacitors or deep trench storage capacitors. The large capacitors are unnecessary since the cell now is active in nature and the transistor provides a large gain. The present invention further provides structures and methods for p-channel floating gate transistors which avoid p-channel threshold voltage shifts and achieve source side tunneling erase

It is to be understood that the above description is intended to be illustrative, and not restrictive. Many other embodiments will be apparent to those of skill in the art upon reviewing the above description. The scope of the invention should, therefore, be determined with reference to the appended claims, along with the full scope of equivalents to which such claims are entitled.